A NONLINEAR ALGORITHM FOR EPOCH MARKING IN SPEECH SIGNALS USING POINCARÉ MAPS

Iain Mann and Steve McLaughlin
Dept. of Electrical Engineering, University of Edinburgh,
King's Buildings, Mayfield Road,
Edinburgh. EH9 3JL. UK
Tel: +44 131 6505655; fax: +44 131 6506554
`e-mail: inm@ee.ed.ac.uk, sml@ee.ed.ac.uk`

ABSTRACT

A novel nonlinear epoch marking algorithm is proposed for use with voiced speech signals. Epoch detection is useful for speech coding, synthesis and recognition purposes, as it provides both the moment of glottal closure and the instantaneous pitch. Our technique functions entirely in state space, by operating on a three dimensional reconstruction of the speech signal which is formed by embedding. By using the fact that one revolution of this reconstructed attractor is equal to one pitch period, we are able to find points which are pitch synchronous by the use of a Poincaré section. Evidently the epoch pulses are pitch synchronous and therefore can be marked. Results using real speech signals are presented to illustrate the performance of the technique.

1 INTRODUCTION

In this paper we describe a novel nonlinear epoch marking algorithm for use with voiced speech. The algorithm makes use of nonlinear dynamical theory by reconstructing the system in state space and then using Poincaré sections to mark pitch synchronous points in the speech signal (*i.e.* the epochs).

Epoch detection has been a pervasive issue in speech processing for many years, since knowledge about the instantaneous pitch or moment of glottal closure is extremely important in speech coding, synthesis and recognition. A number of algorithms for epoch determination exist, most of which operate on the time domain speech signal. Various measures are employed to locate the epoch pulses, such as maximum–likelihood detection [1], discontinuities in the LPC residual [2], similarity models [3] and dynamic programming [4].

The new algorithm reported here should not be taken as a competitor to these existing techniques. Rather, it is a demonstration of the practical possibilities that nonlinear signal processing has to offer in the field of speech processing. Speech has been recognised as being a nonlinear process for a number of years [5], but this offers little advantage if we are unable to exploit this new knowledge. Researchers have already shown the possibilities available for speech coding (*e.g.* [6]); here we propose another direction to pursue. We will firstly review state space reconstruction and the theory of Poincaré sections. Next the algorithm is presented, followed by some results of epoch marking on real speech signals. We conclude with a discussion of the problems and possibilities found with this new technique.

2 STATE SPACE AND THE POINCARÉ SECTION

In nonlinear processing a d–dimensional system can be reconstructed in an m–dimensional state space from a single dimension time series by a process called embedding. Takens' theorem states that $m \geq 2d + 1$ for an adequate reconstruction [7], although in practice it is often possible to reduce m. Time delay embedding involves forming a state space trajectory matrix $\mathbf{X}$ by passing a window of length m through the time series $\mathbf{x}$:

$$\mathbf{X} = \begin{pmatrix} x_0 & x_{\tau_d} & \cdots & x_{(m-1)\tau_d} \\ x_1 & x_{(1+\tau_d)} & \cdots & x_{(1+m\tau_d)} \\ x_2 & x_{(2+\tau_d)} & \cdots & x_{(2+(m+1)\tau_d)} \\ \vdots & & & \vdots \end{pmatrix} \qquad (1)$$

where τ_d is a delay time chosen so as to optimally open up the attractor. An alternative is singular value decomposition (SVD) embedding [8], which may be more attractive in real systems where noise is an issue. This technique partitions the state space into two subspaces, one containing the signal and the other the noise. The singular value decomposition of $\mathbf{X}$ is given by:

$$\mathbf{X} = \mathbf{U}\mathbf{W}\mathbf{V}^{\mathrm{T}} \qquad (2)$$

where $\mathbf{W}$ is diagonal, containing the singular values $w_0 > w_1 > w_2 > \ldots > w_{p-1} \geq 0$ (p being the SVD window length) and $\mathbf{U}$ and $\mathbf{V}$ are orthogonal and contain the associated singular vectors. A reduced trajectory matrix can be formed by:

$$\mathbf{X}' = \mathbf{X}\mathbf{V}_d \qquad (3)$$

where $\mathbf{V}_d$ only contains the columns of $\mathbf{V}$ corresponding to the significant values of $\mathbf{W}$. This process reduces the dimension and removes noise effects [9], leaving a low dimensional, noise–free, attractor reconstruction.

It has been found that vowel sounds are low dimensional, and can be modelled in a 3 dimensional state space [10]. Figure 1 shows an example of time delay and SVD embedding of

This work was supported by an EPSRC CASE award with BT Labs, Martlesham. SML also acknowledges the support of the Royal Society.

the vowel /i/ in 3D state space, illustrating the smoothing capabilities of the SVD process. A time delay of 10 samples or SVD window length of 50 samples has been found to be adequate to ensure that the attractor is opened up at a sampling rate of 22kHz. This reconstruction is pitch synchronous in that one revolution of the attractor is equivalent to one pitch period. Clearly this fact can be exploited to mark points in time separated by multiples of the pitch period. The key to this in practice is the use of a Poincaré map. A Poincaré map replaces the flow of an n–th order continuous system with an $(n-1)$–th order discrete time system [11]. Thus in the case of voiced speech, the 3 dimensional attractor is replaced by a 2 dimensional cross–section. Figure 2 shows this principle, where the hyper–plane Σ cuts the flow, $\mathbf{h}$, of the attractor normally at a chosen point. Crossings in one direction only (*e.g.* from Σ^- to Σ^+) result in a one–sided map, whereas the entire set of crossings produces a two–sided map.

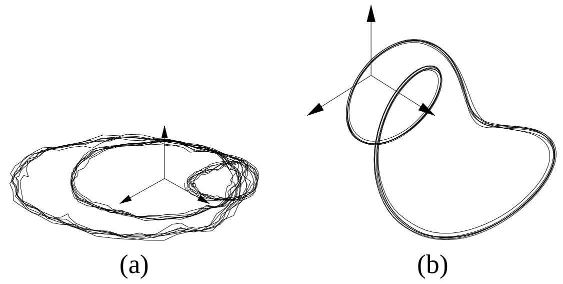
(a) (b)

Figure 1: *(a) Time delay ($\tau_d = 10$ samples) and (b) SVD ($p = 50$ samples) embedding for the vowel /i/. $F_s = 22$kHz in the original signal.*

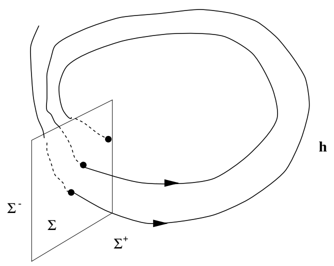

Figure 2: *An example of a (one–sided) Poincaré map, showing the hyper–plane Σ cutting the flow $\mathbf{h}$ normally.*

3 THE NEW EPOCH–MARKING ALGORITHM

Our algorithm uses the principle outlined above to mark successive epochs. To cope with the inherent non-stationarity of the speech signal, it operates on a frame–by–frame basis. Given that the input waveform is a voiced sound (no voiced/unvoiced decision module is included in the prototype), frames of length N samples are read and processed, with a 50% overlap between adjacent frames. In our experiments we have used a frame length of 35msecs ($N = 770$ at 22kHz), since speech is often assumed to be stationary for 30 to 45msec. Each frame is then embedded into 3D state space using an SVD window length of 50 samples. Assuming that the location of the very first epoch in the waveform is known, and denoting that point as x_{GCI_0}, a Poincaré section Σ is positioned normally to this point and all the points $\mathbf{x}_{CROSS}$ $(0 \leq x < N)$ that traverse Σ are found. $\mathbf{x}_{CROSS}$ will contain all of the epochs, since they are pitch synchronous with x_{GCI_0}, but other arbitrary points will also be present, dependent upon the shape and complexity of the attractor.

To chose the correct points from $\mathbf{x}_{CROSS}$, we employ a distance measure, $\langle d_m \rangle$, which locates the points closest to x_{GCI_0} in state space. This distance measure tracks an intersect point x_{CROSS} as it moves around one revolution of the attractor, and finds the average distance in state space of that point from the corresponding movement of x_{GCI_0}:

$$\langle d_m \rangle = \frac{1}{R} \sum_{i=0}^{R-1} \mathrm{D}\left(x_{(GCI_0+ti)}, x_{(CROSS+ti)}\right) \qquad (4)$$

where $D(a, b)$ is the Euclidean distance between points a and b in 3D state space, and $t = (T_0 F_s / R)$ is $1/R$–th of a revolution around the attractor (in samples) at the local point, with local pitch period T_0 seconds. R is typically 8 or 10.

These points should be those within the same part of the attractor manifold as x_{GCI_0}, and hence will be the epoch points. However, due to the complexity of the attractor for certain voiced sounds, other parts of the attractor containing intersections with Σ may also pass close to x_{GCI_0}. To prevent these points accidentally being chosen we use a speech–specific windowing measure similar to that used in [12]. The average pitch period $\langle T_0 \rangle$ is initially calculated using the autocorrelation method with centre clipping over at least 30msec of the frame, and is then updated as each new epoch is marked. We then search in a window between $0.6 \langle T_0 \rangle$ and $1.4 \langle T_0 \rangle$ onwards from the previously marked epoch for the closest point to x_{GCI_0} in state space, since the pitch is not expected to vary by more than $\pm 40\%$ within a voiced section [13]. The new point found is marked as an epoch and the process continues through the frame. Once a frame has been processed, the last epoch marked is taken as x_{GCI_0} for the next frame, and so on through the data. The flow chart shown in Figure 3 illustrates this process.

4 RESULTS

Initially the algorithm was tested using constant pitch vowels from our own database, which also includes laryngograph data for comparative purposes. Further tests where made using a series of rising pitch vowels provided by BT Labs[1], and then signals from the Keele University Pitch Extraction Database [14], which provides speech and laryngograph data from 15 speakers reading phonetically balanced sentences. In all cases the sampling rate used was 22kHz so as to adequately fill the state space reconstruction (our database was recorded at 22kHz; the other signals were up–sampled, the BT vowels originally being at 12kHz and the Keele signals at 20kHz). All the signals have 16 bit resolution.

[1] Thanks to A. Lowry for providing this data.

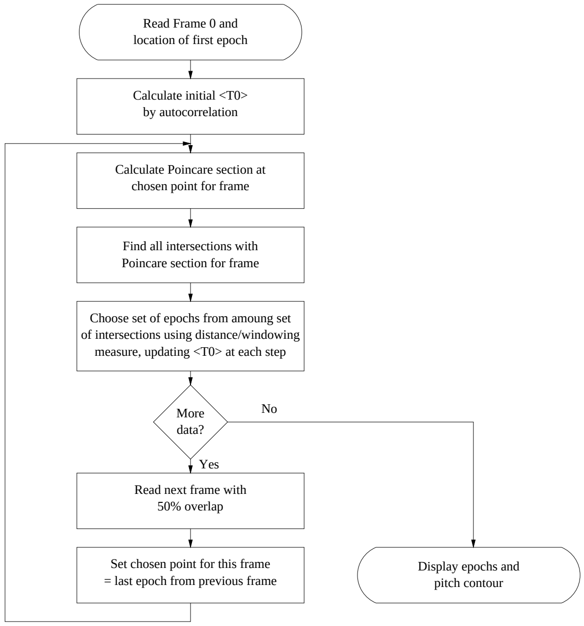
Read Frame 0 and
location of first epoch
Calculate initial <T0>
by autocorrelation
Calculate Poincare section at
chosen point for frame
Find all intersections with
Poincare section for frame
Choose set of epochs from amoung set
of intersections using distance/windowing
measure, updating <T0> at each step
More
data?
No
Yes
Read next frame with
50% overlap
Set chosen point for this frame
= last epoch from previous frame
Display epochs and
pitch contour

Figure 3: *Schematic of the epoch marking algorithm.*

Figure 4 shows a snapshot of the algorithm during processing on the constant pitch, stationary vowel /a/ spoken by a female speaker. The epochs are marked as calculated on the time domain waveform. The left–hand–most epoch is that which was used as x_{GCI_0}. The attractor reconstruction shows how the Poincaré section, which intersects the flow at this point, crosses through several other parts of the manifold. This is seen more clearly on the 2–sided Poincaré map. In this case the attractor structure is relatively simple resulting in four sets of intersection points, all of which are well separated making classification of the points easy. The points corresponding to the epochs are those enclosed within the circle.

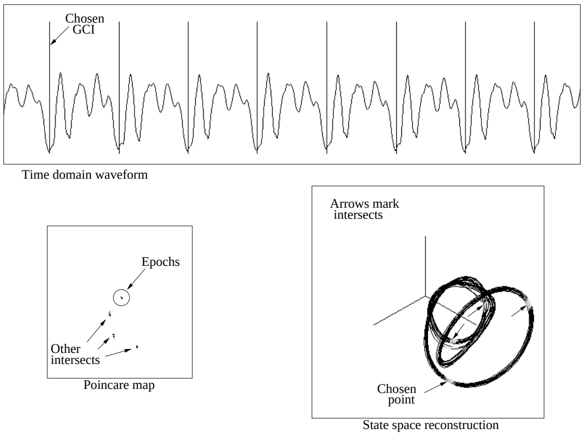

Figure 4: *Snapshot of processing of the stationary vowel /a/, showing a frame of the signal with the calculated epoch markers (top); the two–sided Poincaré map of intersections with the attractor (bottom left); the attractor reconstruction with intersects indicated (bottom right).*

The algorithm was found to correctly mark all of the stationary vowels considered.

As a next step we considered a simple non–stationary case, that of rising pitch vowels. Figure 5 shows the results of our algorithm compared to those produced by a dynamic programming–based approach [4][2] for the vowel /u/, pronounced by a male speaker. Again the left–hand–most epoch is that which was marked as x_{GCI_0}. It can be seen that the performance of our new algorithm is equal to that of this established approach, demonstrating that the frame–by–frame operation allows us to track changes in the attractor structure, caused in this case by the changes in pitch.

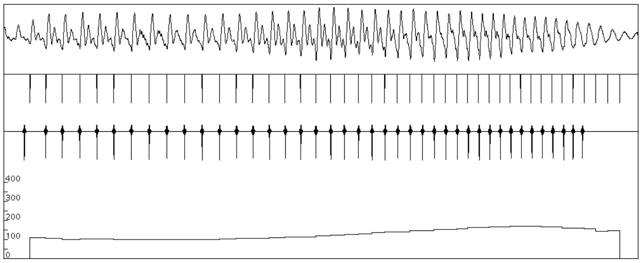

Figure 5: *Results for the rising pitch vowel /u/. From top to bottom: the signal; the epochs as calculated by our algorithm; the epochs as calculated by the dynamic programming approach; the pitch contour (Hz) resulting from our algorithm.*

Finally we tested the algorithm with various voiced segments from the Keele University database. Figure 6 shows the algorithm's performance on a section of the phrase "the northwind and the sun were disputing", spoken by a male speaker. There is a considerable change in the signal, and hence in the attractor structure, in this example, yet the epochs are still mostly well located when compared against the laryngograph signal. However problems have been encountered with other signals tested, usually caused by the state space reconstruction being very complicated, thus making the selection of the correct intersect points difficult. In general, the more complicated state space structures occur at low pitch values (many oscillations within one pitch period) and with voiced sounds where fricative noise also occurs (which tend to fill the state space).

5 DISCUSSION

Clearly a major drawback to this algorithm is the need to know the location of the first epoch. Due to the nature of the technique, it is only possible to mark points which are pitch synchronous; we cannot tell *a priori* if a data point is a glottal closure instant, but once one is known all subsequent epochs in that voiced section can, in theory, be found. Unfortunately the attractor's geometric structure does not provide any information about the epoch locations either, so at present this problem is unresolved.

[2] Available commercially as Entropic Research Laboratory's ESPS epoch function.

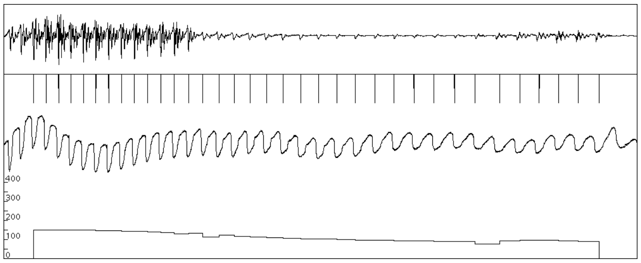

Figure 6: *Results for the voiced section of "sun were" from the Keele database. From top to bottom: the signal; the epochs as calculated by our algorithm; the laryngograph signal; the pitch contour (Hz) resulting from our algorithm.*

This aside, the technique appears useful. It works very well on the simple cases of stationary vowels and rising pitch vowels, accurately marking all the epochs. When applied to real speech signals we have met with moderate success. The algorithm is often able to track quite considerable changes in the attractor structure, caused by changes in vowel sound and/or pitch, but should an error occur (usually caused by a misalignment of the Poincaré section, often when the attractor structure is very complicated) it is unable to recover leading to incorrect marking at all points forward from the error point. This is again due to the fact that we are not actually locating the epoch pulses specifically, only points which are pitch synchronous. Therefore when an error occurs, causing a loss of synchronisation, it propagates through the remainder of the signal. Further work will need to address this problem, as well as improving the algorithm so it is better able to cope with sudden changes in attractor structure.

6 CONCLUSIONS

We have shown how nonlinear signal processing theory can be applied to the practical problem of epoch marking. The proposed algorithm, which operates in state space using a Poincaré section to mark pitch synchronous points, has been shown to perform very well on simple vowel sounds and to give promising results on real voiced speech signals. Therefore this novel technique demonstrates the potential of nonlinear theory in speech processing.

7 ACKNOWLEDGEMENTS

The authors would like to thank G. Kubin of the University of Vienna and M. Edgington of BT Labs for useful discussions about this work.

References

[1] Y. M. Cheng and D. O'Shaughnessy, "Automatic and reliable estimation of glottal closure instant and period," *IEEE Transactions on Audio, Speech and Signal Processing*, vol. 37, pp. 1805 – 1815, December 1989.

[2] R. J. DiFrancesco and E. Moulines, "Detection of glottal closure by jumps in the statistical properties of the speech signal," *Speech Communication*, vol. 9, pp. 401 – 418, December 1990.

[3] Y. Medan, E. Yair, and D. Chazan, "Super resolution pitch determination of speech signals," *IEEE Transactions on Signal Processing*, vol. 39, pp. 40 – 48, January 1991.

[4] D. Talkin, "Voicing epoch determination with dynamic programming," *Journal of the Acoustical Society of America*, vol. 85, Supplement 1, p. S149, 1989.

[5] G. Kubin, *Speech Coding and Synthesis*, ch. Nonlinear Processing of Speech, pp. 557 – 610. Elsevier, 1995.

[6] B. Townshend, "Nonlinear prediction of speech," in *International Conference on Acoustics, Speech and Signal Processing*, vol. 1, pp. 425 – 428, 1991.

[7] F. Takens, "Detecting strange attractors in turbulence," in *Proceedings of Symposium on Dynamical Systems and Turbulence* (A. Dold and B. Eckmann, eds.), pp. 366 – 381, Lecture Notes in Mathematics, 1980.

[8] D. S. Broomhead and G. P. King, *Nonlinear Phenonema and Chaos*, ch. On the Qualitative Analysis of Experimental Dynamical Systems, pp. 113 – 144. Bristol: Adam Hilger, 1986.

[9] A. I. Mees, P. E. Rapp, and L. S. Jennings, "Singular–value decomposition and embedding dimension," *Physical Review A*, vol. 36, pp. 340 – 346, July 1987.

[10] A. Kumar and S. K. Mullick, "Nonlinear dynamical aspects of speech," *Journal of the Acoustical Society of America*, vol. 100, pp. 737 – 793, September 1996.

[11] T. S. Parker and L. O. Chua, *Practical Numerical Algorithms for Chaotic Systems*. New York: Springer–Verlag, 1989.

[12] C. Murgia, I. Mann, and G. Feng, "An algorithm for the estimation of glottal closure instants using the sequential detection of abrupt changes in speech signals," in *European Signal Processing Conference*, (Edinburgh), pp. 1685 – 1688, 1994.

[13] W. Hess, *Pitch Determination of Speech Signals*. Springer–Verlag, 1983.

[14] F. Plante, G. F. Meyer, and W. A. Ainsworth, "A pitch extraction reference database," in *EUROSPEECH'95*, vol. 1, pp. 837 – 840, September 1995.